SANDIA REPORT
SAND2014-16084
Unlimited Release
Printed July 2014

Genomic Functionalization: The Next Revolution In Biology

Joe S. Schoeniger, Peter Anderson, Paula M. Imbro

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online order:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2014-16084
Unlimited Release
Printed July 2014

Genomic Functionalization: The Next Revolution In Biology

Predicting Protein Function From Molecular Interactions

Peter Anderson, Joe S. Schoeniger, Paula M. Imbro,
Biological & Materials Science and Systems Biology
Sandia National Laboratories
P.O. Box 969 MS9292
Livermore, CA 94551

Abstract

We have implemented a ligand-alignment algorithm into our developed computational pipeline for identifying specificity-determining features (SDFs) in protein-ligand complexes. Given a set of protein-ligand complex structures, the algorithm aligns the complexes by ligand rather than by the Cα-RMSD or standard approach, providing a single reference frame for extracting SDFs. We anticipate that this ligand-alignment capability will be highly useful for protein function prediction. We already have a database containing > 20 K ligand-protein complex crystal structures taken from the Protein Data Bank. By aligning these proteins to single reference frames using ligand alignment, we can submit the complexes to our pipeline for SDF extraction. The SDFs derived from this training procedure can be used as thumbprints that are hallmarks of individual enzyme classes. These SDF thumbprints may then serve as guides to the prediction of function of new unknown proteins.

PAGE INTENTIONALLY LEFT BLANK

CONTENTS

FIGURES

TABLES

NOMENCLATURE

DNA	deoxyribonucleic acid
HIV	human immune virus
HMDB	human metabolome database
HMM	hidden Markov models
mRNA	messenger RNA
NGS	next generation sequencing
RAST	rapid annotation using subsystem technology
RNA	ribonucleic acid

PAGE INTENTIONALLY LEFT BLANK

1. INTRODUCTION

Predicting the phenotype of an organism from genetic information is a fundamental goal of genetics. As defined in the neo-Darwinian Synthesis during the early and mid part of the twentieth century by Fisher, Haldane, Huxley and others, the association between genes and phenotypes was conceptualized as a probabilistic problem, and an objective was the description of probabilities of trait distributions in breeding populations, by way of predicting and understanding the genetic underpinnings of evolution. An abundance of morphological, clinical, behavioral and, later, biochemical traits could be measured across a population, and related, more or less indirectly (and plausibly) to fitness and evolutionary selection. The accuracy of phenotype predictions was, however, limited in a number of ways: Epigenetic, physiologic, environmental or developmental factors modulate the expression of genes and result in trait variability. Many traits are polygenenic. The resulting effect (i.e., multifactorial inheritance with variable penetrance) made it difficult to quantify many traits, and created a grey zone between complex traits and merely complicated observable features of organisms (especially morphological and developmental features). Most fundamentally, however, there was very limited knowledge of the nature of genes, in contract to the very large number of readily observable phenotypes.

Typically, the genetic profile of a parent population was inferred from prior histories of breeding and trait distribution. There was no general way of physically determining the genetic composition of an organism, or even the particular catalog of alleles present in a given individual until the advent of gene sequencing. (For example, genome sequencing has revealed that, due to variable gene duplication, there are large copy number variations in different alleles between different individuals). Breakthroughs such as amino acid sequencing of proteins and defining the genetic code allowed biochemical phenotypes of enzymes to be correlated with particular genotypes. With the establishment of the central dogma of molecular biology and creation of effective DNA sequencing methods, a comfortable equilibrium was established in which genes were identified with single protein sequences which were identified with single molecular functions which depended on the sequence and structure of the protein. The combination of cloning technologies and gene sequencing meant that protein sequences could be more easily obtained, and molecular mechanisms underlying a variety of genetic phenomena explored. In general, this was carried out in a highly directed context: a particular biological function was studied, and a chromosomal locus was identified using genetic methods, or via a sequence tag from the protein of interest, or based on interesting messenger RNA.

The first reconstructions of the sequences of the entire genomes of complex organisms were accomplished using well-established methods for cloning and gel-based sequencing of short pieces of DNA, deployed at a massive scale. The concept for what a genome actually was at the time, might have been described as collection of protein coding genes and their local regulatory elements, plus a small number of genes encoding (mainly structural) RNA, and other miscellaneous features, strung out on a chromosome, which was revealed, at least in higher eukaryotes, to be 99% "junk". Although sequencing the human genome was a momentous scientific achievement, this gene-centric focus lead to astonishment at the small number of genes discovered, compared to the gene count of other organisms "less complex" than humans.

With the completion of every new genome, predicting the function of apparent protein-coding genes revealed by genomic sequencing became a major task (often tackled by crowds of experts at "annotation parties"). Putative functions were assigned mainly based on gene sequence homology to proteins with known functions. As techniques for determining protein structure improved, it also became possible to compare the structures of new and existing protein, and to try to infer the function of one from that of the other. Biochemical and genetic validation of the annotations of genes in new organisms is difficult- in fact it is almost never done on a systematic, genome-wide basis. Furthermore, whole genome sequences, particularly of bacteria, also reveal a large number of features that look like genes but code for proteins with no sequence or structural homology to known proteins. A significant fraction, in the tens of percent, of a genome may be composed of these so-called "hypothetical" proteins. Even if cloned, produced in a heterologous expression system, and purified, it is difficult to understand what biochemical experiments should be pursued to predict the function of such a protein. As described below, obtaining or predicting the structure of such a protein may or may not be much help, other than to show its structural homology to other hypothetical proteins.

Prior to, or contemporaneously with, the first major genome sequencing efforts, advances in molecular biology also lead to discoveries such as ubiquitous mRNA splicing in higher eukaryotes, and numerous gene regulatory mechanisms (often involving RNAs). Biochemical and structural biology advances, especially in cell signaling, also made it clear that many proteins interact with multiple partners, engage in multiple biochemical pathways. Pleitropic effects were therefor potentially very widespread, and the notion of "one gene, one function" was shown to be of limited veracity and utility. When one gene influences multiple phenotypic traits, it may do so in combination with several other different pleitropic genes, forming pleitropic networks that make the assignment of function even more fuzzy. In other words, the both the notions "gene" and "function" have become harder to define, and it may be impossible to assign "functions" to "genes" on a one-to-one basis.

Next generation sequencing (NGS) technologies have enabled revolutionary advances in areas such as detection and diagnosis of biological agents for human health and biodefense, forensics, and bioenergy. These approaches have also led to an exponential growth in the rate of production of biological sequence data. The principal challenge in analyzing genomic sequence data is still predicting the functional significance of most genome subsequences, or genomic functionalization. The knowledge gap between gene sequence and function complicates tasks such as assessing the potential pathogenicity of a biological agent, understanding the health implications of human genomic sequence variations, prospecting for new enzymatic activities capable of catalyzing novel chemical transformations, or understanding how sequence changes in existing enzymes may affect the function of enzymes involved in processes such as cellulosic biomass deconstruction.

In a larger context, NGS, and the ever-increasing number of available sequenced genomes, has created an inverse problem to the one that confronted geneticists in the mid twentieth century. Instead of a large amount of observable phenotypic information and a limited amount of murky genetic information, we now have a superabundance of genetic information, in its genomic context, and a relative dearth of phenotypic information, while notions of phenotype and function become ever more complex. Methods are needed to computationally predict the function of new protein sequences that go beyond comparison of sequence homology or prediction of structural homology to known proteins. These methods should be as universally applicable as possible, and help build a bridge to high-throughput experimental approaches for evaluating biochemical function. Finally, they should accommodate the notion that proteins may have multiple biological roles.

2. FUNCTION PREDICTION AT THE MOLECULAR LEVEL

Difficulties with predicting gene function do not, of course, originate uniquely within the recent context of interpreting next generation sequencing data or analysis of large numbers of whole genomes. The problematic nature of genes and the conceptual complexities of biological function not withstanding, it is still important to understand and predict the biochemical function of individual proteins. Without knowledge of these molecular functions, understanding of the functioning of cells and organisms remains incomplete. From a biochemical and physiological perspective, we cannot apply reductive causal analyses of biological systems if we do not know what the parts do. From a genetic and evolutionary perspective, gene frequencies, epistasis, and co-expression patterns cannot be causally understood if connections cannot be made between the biochemical function of different genes. At the very least, a gene that has no function is expected, over time, to accumulate mutations which will eventually result in it not being expressed, or lost altogether. The gradual conversion of a gene into a pseudogene generally results in readily discernable sequence changes, so intact genes generally have at least one valid function, unless this function has been lost very recently in evolution.

The notion of function is still complex, even when narrowed to a single protein-producing gene. For example, gene ontologies attempt to describe a bundle of gene attributes, such as interaction partners, pathway memberships, cellular compartments, etc. Many of these attributes are evolutionarily conserved. Features such as the cell compartment a protein is found in may also be associated with gene features such as signal sequences. More generally, there are a number of nucleic acid or amino acid sequence motifs that have strong functional implications, including signal sequences, glycolsylation sites, metal binding sites, etc. In general, though, features such as these delimit attributes such as cell compartmentalization, but do not positively predict a biochemical role.

Classically, the most important biochemical role of proteins is that of enzymes. Solving the narrow problem of predicting enzymatic function is important to many applications in biodefense, public health, and bioenergy. Enzymes, in particular, are the targets of bioprospecting for new catalysts, and many pathogenesis determinants are enzymes (e.g., many drug resistance genes, toxins and effectors).

Enzyme function prediction is, however, very challenging, for several reasons. Even within the context of a finished genome, despite the existence of high-quality gene prediction and annotation tools such as RAST (REF), genes are primarily annotated based upon their sequence homology to “known” genes and operons. It may be difficult to even find genes shorter than 50-100 amino acids in a new genome due to the limited length of sequence available for comparison. But even more basically, assigning the correct enzymatic function to a new sequence based on sequence or structural homology to a known gene depends on 1) the function of the known gene having been correctly determined, and 2) assumes that the homology implies conservation of the function. With regard to the first assumption, there are relatively few model organisms that have been thoroughly biochemically characterized. It is laborious to conduct genetic surveys of biochemical phenotypes and to purify and express proteins in order to characterize them chemically and piece together metabolic pathways. The most closely homologous gene to a new gene is not necessary one from a model organism, and may not be a gene for which the function has ever been experimentally verified. This leads to a daisy chain of secondary, tertiary, etc. annotations for each newly discovered gene. Leaving aside the question of whether the original annotation was correct, there is a basic problem that functions change based on the environment of the organism. Genes annotated as having a role in pathogenesis in one organism, may, for example, be present in apathogenic relatives (e.g., type III secretion components).

The second assumption, that the homology implies conservation of the function, is in turn contingent on the idea that protein sequence determines structure and that structure determines function. In general, enzymatic function is determined by the active site structure, not by the fold of a protein. Convergent evolution may result in quite diverse protein backbone structures providing a scaffold for essentially identical active sites. Conversely, active site structure may rapidly evolve with few changes in fold or backbone structure, in the presence of strong selection pressures or in rapidly mutating organisms such as viruses. Finally, many novel metabolic and pathogenic genes are present in genomic islands (often the result of phage integration into the genome or plasmid) and these islands have a high incidence of so-called "hypothetical genes". These genes may or may not be truly hypothetical, in that many hypothetical proteins are known to be expressed, based on proteomic surveys, but the function of a hypothetical gene is generally mysterious, and even strong homology of a new gene to a known hypothetical does not inform function.

NGS has compounded the difficulty of gene function prediction in at least three ways. First, it has simply increased the rate of production of new genomes, and of mobile gene elements such as plasmids. Second, the data may not be from a single organism, and data may be in the form of random pieces of genomic sequence or gene transcripts from a so-called metagenomic sample containing many diverse microbial species. This data may or may not provide sufficient coverage to allow for assembly into full genes, though technology improvements mean that the current length of a sequenced read is around 500bp and improving. Third, the combination of single cell DNA/RNA amplification methods and NGS allows for the recovery of significantly portions of the genomes of unculturable organisms. These organisms may have considerable sequence divergence from previously characterized organisms.

Specifically, we are attacking the important general problem of predicting the functions of protein based on their structures and binding interactions. Our approach is to create methods for extracting and classifying structural and chemical features from the enzyme binding sites that determine their binding specificity, across large fields of diverse ligands and across diverse protein structures. These features represent binding motifs that are solutions to molecular recognition reflecting convergent evolution across diverse life forms. As such they may serve as fairly universal functional classifiers, and form the basis for sequence patterns that are detectable without structural modeling.

3. FUNCTION BASED ON SMALL MOLECULE BINDING

Our approach for predicting function relies on the tie between protein function and the specific binding of that protein to certain types of other molecules (a ligand test set). Extracting the features of binding interfaces of proteins that predict the specificity of their binding interactions enables supervised classification of proteins, using these feature sets, according to their relative differences in binding site chemistry and specificity. To enable testing, we restricted the diversity of proteins targeted to have an unambiguous important binding site (enzymes). This enabled successful prediction of E.C. classes. Geometric alignment of features of highly diverse proteins was also problematic. We have developed an algorithm that aligns the complexes by ligand rather than by protein Cα-RMSD, which should enable classification of a broader range of proteins targets.

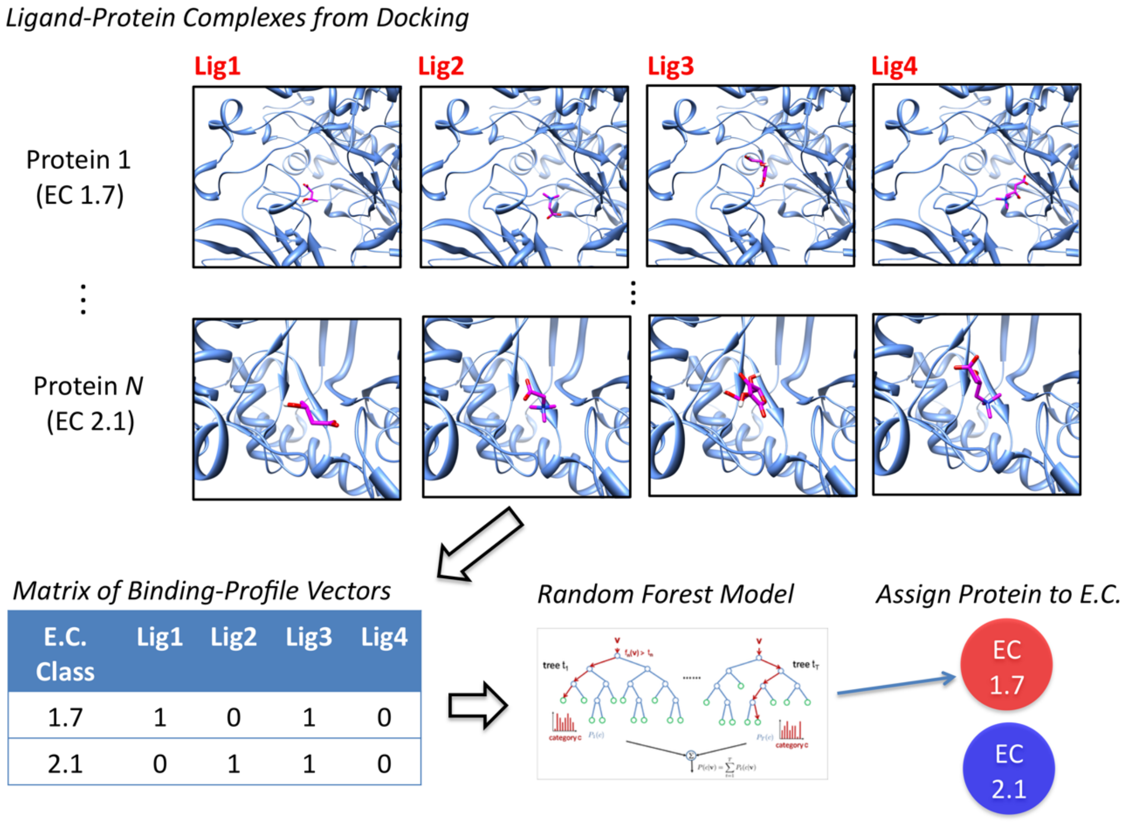
We have completed computational docking of a panel of 100 common metabolites to the active sites of representative enzymes from ~800 unique enzyme classification (E.C.) numbers. In total, docking has been performed against ~10,000 enzymes, representing ~1 million modeled enzyme-ligand complexes. Metabolites were selected from the HMDB Serum Metabolome (www.serummetabolome.ca) on the basis of serum concentrations and subject to the following two constraints: (i) molecular weight > 100 Daltons; (ii) molecules are organic not inorganic. The 100 most concentrated metabolites satisfying both constraints were selected for modeling.

Our ultimate goal is to predict the functions of unknown proteins from their sequences using a computational pipeline we have recently developed [1]. This pipeline extracts sets of specificity-determining features (SDFs) for clusters of proteins or ligands that have similar binding properties to one another. The extracted SDFs are predictive of membership within a given binding clade and provide easily visualized models of the structural features of proteins/ligands that underlie their shared binding profiles with others proteins/ligands. We hypothesize that it is possible to develop a database of SDFs that are characteristic of different classes of proteins with previously characterized function, for instance, enzymes within a given E.C. class. We further hypothesize that not only can we develop such a database of SDFs, but also that computational techniques such as Hidden Markov Models (HMM) can link protein sequences to specific SDF sets. In principle, then, we could use HMM methods to predict the SDF set of an uncharacterized protein based on the protein's amino acid sequence. The predicted SDF set would allow the protein to be assigned to a certain functional class within the SDF database.

A major challenge in accomplishing this task is to surmount a technical hurdle inherent in our computational pipeline as it now stands: the alignment of unrelated proteins' active sites to a common reference frame. Given that our pipeline was developed to analyze the binding of small molecules to proteins within the same family (e.g. human protein kinases and HIV-1 protease sequence variants) and that these proteins are easily aligned, the pipeline uses a single universal coordinate set to calculate and describe SDFs. For this reason, it cannot be used for proteins that have poor alignments, e.g. for proteins from different enzyme classes. We are currently implementing methods to overcome this limitation, including the encoding of ligand-protein interactions as 1D fingerprints that are independent of protein alignment and applicable to any ligand-protein complex [2].

An easily tractable and useful first step in leveraging the large volume of modeling data that we have obtained so far is to predict the function of proteins based on whether they do or do not bind to a standard set of small molecules. A recent paper shows that proteins that experimentally bind with similar profiles to a panel of ligands have similar functions [3]. More specifically, the study shows that protein pairs whose binding profiles (encoded as binary vectors where "1" and "0" represent ligand binding and nonbinding, respectively) have a low E-value relative to one another demonstrate a high functional similarity score based on GO terms. In contrast, protein pairs having binding profiles with higher E-values show lower functional similarity scores. In addition to being fast and amenable to high-throughput experiments, predicting protein function based on ligand binding profiles could provide useful first clues about SDFs themselves. Previous structural and bioinformatics studies have shown, for example, that proteins with similar binding profiles frequently have structurally similar active site sub-cavities and active site topologies [2,4-6]. Therefore, we believe that computationally analyzing similarities in protein binding patterns based on our modeling data can take us one step closer to both protein function prediction and SDF determination.

In order to test the functional predictivity of binding profiles, we randomly selected 89 enzymes from two distinct enzyme class/subclass groups (first two E.C. numbers). Specifically, we selected 42 enzymes from E.C. 2.1 (transferases that transfer one-carbon groups) and 47 enzymes from E.C. 1.7 (oxidoreductases acting on nitrogenous compounds as donors). Using the *in silico* scoring function XScore [7], enzyme-ligand binding affinities (expressed as dissociation constants, K_d) were predicted for each enzyme interacting with each of the 100 serum metabolites based on their docked binding conformations. As shown in Figure 1, each enzyme's binding profile was encoded as a binary vector consisting of 89 entries, where "1" indicates a predicted $\log(K_d) \geq 4.5$ (ligand binds) and "0" indicates a predicted $\log(K_d) < 4.5$ (ligand does not bind). Each binding profile vector also contained a classification label for the E.C. class/subclass. The 89x100 matrix of computational binding profiles was provided as a training set for induction of a random forest machine-learning model to be used for E.C. classification and prediction (Figure 1).



E.C. Class	Lig1	Lig2	Lig3	Lig4
1.7	1	0	1	0
2.1	0	1	1	0

Figure 1. Ligand-Protein Complexes from Docking

Accomplished workflow for predicting enzyme function based on metabolite binding profiles. A panel of 100 metabolites was docked to ~10,000 enzymes. Randomly selected enzymes from two E.C. class/subclass designations had their metabolite binding profiles encoded as binary vectors. These vectors were used to train a random forest model for classification of new proteins. The model assigns 73 out of 89 enzymes to their correct E.C. classifications.

The random forest model created from the training data predicts E.C. classification with good accuracy. The overall out-of-bag error rate is 18.0%. A confusion matrix for the out-of-bag classification is shown in Table 1.

Table 1. Confusion matrix for classification of E.C. numbers based on computational metabolite binding profiles. Binding profiles are binary vectors.

	E.C. 1.7	**E.C. 2.1**	**Classification Error**
E.C. 1.7	41	6	12.8%
E.C. 2.1	10	32	23.8%

As shown in Table 1, the random forest model assigns 73 of the 89 proteins to their correct E.C. class. The use of simple binary binding-pattern vectors for enzyme binding to 100 metabolites is sufficient, based on these findings, to allow accurate classification of enzyme function. Studies are ongoing to verify that this classification accuracy is reproducible for other pairs of E.C. class/subclass designations and for pairs of higher-level E.C. class pairs, e.g. a set of randomly selected oxidoreductases versus a set of randomly selected transferases (E.C. 1 vs E.C. 2).

We conducted a similar random forest classification test where the protein binding-profile vectors consist not of simple "1" and "0" entries, but instead contain the actual predicted ligand K_d values as entries. The trained random forest model for this data set performs even better than the model for binary vectors, achieving an overall out-of-bag error of 13.5%, as shown in the confusion matrix of Table 2. The random forest model accurately assigns 77 of the 89 proteins to their correct E.C. class.

Table 2. Confusion matrix for classification of E.C. numbers based on computational metabolite binding profiles. Binding profiles consist of predicted ligand Kd values

	E.C. 1.7	**E.C. 2.1**	**Classification Error**
E.C. 1.7	45	2	4.3%
E.C. 2.1	10	32	23.8%

Our reported results are significant for two reasons. First, to our knowledge, this is the first demonstration that enzyme functions can be accurately assigned based on their binding profiles to standard metabolites as predicted by virtual screening. Second, the binding profiles themselves can be clustered in order to yield insight into SDFs. These findings thus have dual utility: (i) we have proof of principle of the feasibility of accurate enzyme function prediction in a sequence-independent, high-throughput fashion, and (ii) the body of data that we generate can itself feed into the next stage of our project, the creation of a SDF database for proteins of diverse functional classes.

4. CONCLUSIONS

We have implemented a ligand-alignment algorithm into our developed computational pipeline for identifying specificity-determining features (SDFs) in protein-ligand complexes. Given a set of protein-ligand complex structures, the algorithm aligns the complexes by ligand rather than by protein Cα-RMSD, which is the standard approach. This function allows us to align any number of complexes regardless of their ligand structures. After a set of complexes has been aligned by ligand structures, we have a single reference frame for extracting SDFs. We anticipate that this ligand-alignment capability will be highly useful for protein function prediction. We already have a database containing > 20 K ligand-protein complex crystal structures taken from the Protein Data Bank. By aligning these proteins--or more manageably sized subsets thereof--to single reference frames using ligand alignment, we can submit the aligned complexes to our pipeline for SDF extraction. The SDFs derived from this training procedure can be used as thumbprints that are hallmarks of, for example, individual E.C. classes. These SDF thumbprints may then serve as guides to the prediction of function of new unknown proteins.

5. REFERENCES

1. Anderson PC, De Sapio, V, Turner KB, Elmer SP, Roe DC, Schoeniger JS; *J. Med. Chem.* 2012, *55*, 1926-1939.

2. Desaphy J, Raimbaud E, Ducrot P, Rognan D: *J. Chem. Inf. Model.* 2013, *53*, 623-637.

3. Shortridge MD, Bokemper M, Copeland JC, Stark JL, Powers R: *J. Proteome Res.* 2011, *10*, 2538-2545.

4. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G: *J. Med. Chem.* 2004, *47*, 550-557.

5. Wallach I, Lilien RH: *Bioinformatics* 2009, *25*, i296-i304.

6. Sheridan RP, Nam K, Maiorov VN, McMasters DR, Cornell WD: *J. Chem. Inf. Model.* 2009, *49*, 1974-1985.

7. Wang R, Lu Y, Wang S: *J. Med. Chem*. 2003, *46*, 2287-2303.

Sandia National Laboratories